ABSTRACT

A method for segmenting a compound word in an unrestricted natural-language input is disclosed. The method comprises receiving a natural-language input consisting of a plurality of characters. Next, a set of probabilistic breakpoints based on a probabilistic breakpoint analysis is constructed in the natural-language input. A plurality of linkable components is identified by traversal of substrings of the natural-language input delimited by the set of probabilistic breakpoints. Finally, a segmented string consisting of a plurality of linkable components spanning the natural-language input is returned. The segmented string can be interpreted as a compound word.